Standardised experiments in mutant mice reveal behavioural similarity on 129S5 and C57BL/6J backgrounds

Standardised experiments in mutant mice reveal behavioural similarity on 129S5 and C57BL/6J backgrounds

Louie N. van de Lagemaat[1,2], Lianne E. Stanford[2], Charles Pettit[2], Douglas J. Strathdee[2,3], Karen E. Strathdee[2], Kathryn A. Elsegood[1,2], David G. Fricker[1,2], Mike D. R. Croning[1,2], Noboru H. Komiyama[1,2] and Seth G. N. Grant[1,2]

[1] Centre for Clinical Brain Sciences, Chancellor's Building, University of Edinburgh, Edinburgh, EH16 4SB, UK

[2] Genes to Cognition Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK

[3] Transgenic Technology Division, CRUK Beatson Institute, Glasgow G61 1BD, UK

Correspondence:
Seth G. N. Grant
Chancellor's Building, University of Edinburgh
49 Little France Crescent
Edinburgh EH16 4SB
United Kingdom
Tel: +44 (0) 131 242 7984
Fax:
Email: seth.grant@ed.ac.uk

Running Title: Strain-mutation interactions in mouse behavior

Keywords: genetic background, strain interaction, mutation-strain interaction, analysis of confounds, sampling effect

Abstract

Behavioural analysis of mice carrying engineered mutations is widely used to identify roles of specific genes in components of the mammalian behavioural repertoire. The

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/gbb.12364

reproducibility and robustness of phenotypic measures has become a concern that undermines the use of mouse genetic models for translational studies. Contributing factors include low individual study power, non-standardised behavioural testing, failure to address confounds and differences in genetic background of mutant mice. We have examined the importance of these factors using a statistically robust approach applied to behavioural data obtained from three mouse mutations on 129S5 and C57BL/6J backgrounds generated in a standardised battery of five behavioural assays. The largest confounding effect was sampling variation, which partially masked the genetic background effect. Our observations suggest that strong interaction of mutation with genetic background in mice in innate and learned behaviours is not necessarily to be expected. We found composite measures of innate and learned behaviour were similarly impacted by mutations across backgrounds. We determined that, for frequently-used group sizes, a single retest of a significant result conforming to the commonly used $p<0.05$ threshold results in a reproducibility of 60% between identical experiments. Reproducibility was reduced in the presence of strain differences. We also identified a p-value threshold that maximized reproducibility of mutant phenotypes across strains. This study illustrates the value of standardized approaches for quantitative assessment of behavioural phenotypes and highlights approaches that may improve the translational value of mouse behavioural studies.

Introduction

Engineered mutations in mice are widely used for assessing the roles of genes and proteins in many biological processes including behaviour. However, experiments are complicated by the issue of reproducibility, failures of which may be due to a number of issues. The first major issue is methodological heterogeneity and lack of standardisation (Crabbe *et al.*, 1999, Kilkenny *et al.*, 2009) and the second issue is differences in mouse genetic background. These issues often arise due to the specific questions the studies are addressing, but they introduce uncertainty about the relevance of the studies for others (Crawley *et al.*, 1997, Logue *et al.*, 1997, Rodgers *et al.*, 2002). A third issue is the presence of confounding factors such as differing animal age (Kilkenny *et al.*, 2009). These factors may be addressed in large, standardised phenotyping datasets using mixed effects models (Karp *et al.*, 2012).

Mouse genetic backgrounds are often chosen for particular experiments based on differences in wildtype behaviour, as reviewed by others (Baker, 2011, Balogh *et al.*, 1999, Crawley, 2008, Crawley *et al.*, 1997, Holmes *et al.*, 2002). The two most common genetic backgrounds are 129 and C57BL/6, which were derived in the 1920s from fancier stocks. Of 27,906 mouse lines listed in a recent database (Blake *et al.*, 2014), 78% were made on either one or both of these backgrounds. Anecdotal evidence suggests that the chosen genetic background may occasionally interact with the engineered mutations being tested (Drapeau *et al.*, 2014, Grant *et al.*, 1992, Holmes *et al.*, 2003, Huang *et al.*, 2013, Kastenberger *et al.*, 2012, Mistry *et al.*, 2014, Morice *et al.*, 2004, Pietropaolo *et al.*, 2011, Popova *et al.*, 2009). While the theoretical underpinnings of this effect are well understood (reviewed in Gerlai, 2001), the impact of genetic background on mutant behavioural phenotypes has not been quantified.

The Genes to Cognition (G2C) Programme (Grant, 2003) implemented a single-centre behaviour pipeline approach with rigorous methodological standardisation. Over 2700 mice were identically tested in a behavioural test battery involving five environments, from which 16 maximally independent (uncorrelated) behaviour variables were carefully chosen for similarity between 129S5 and C57BL/6J wildtypes. Amongst 60 G2C mutant lines, the six lines analysed here carried knockout mutations in *Dlg4* (PSD-95), *Dlg3* (SAP102), and *Dusp6*, each back-crossed into both 129S5 and C57BL/6J genetic backgrounds. These

genes encode proteins in the postsynaptic terminal of excitatory synapses and are known to play a role in cognition through behavioural experiments using mazes and computerized touchscreens (Cuthbert *et al.*, 2007, Migaud *et al.*, 1998, Nithianantharajah *et al.*, 2013, Park *et al.*, 2011).

In this report, we apply robust statistical analysis of confounds, including sampling variation, to assess the similarity of mutant phenotypes between background strains. Based on 48 behavioural comparisons, we examined how frequently the behaviours differed across backgrounds. We provide, to our knowledge, the first robust estimate of how often mouse genetic background has a significant modulating effect on mutant behavioural phenotypes, in both innate and learned behaviours. We also identify composite measures of innate and learned behaviour that measure aggregate impacts of mutation and show they are similar across genetic backgrounds. By demonstrating strong similarity of mutant behavioural phenotypes between mouse genetic backgrounds, these analyses provide insight into, and validate comparisons of, mutant phenotypes across studies in rodent model systems. Finally, we developed a p-value threshold approach that shows increased reproducibility of more-significant mutant phenotypes across backgrounds. Taken together, our results demonstrate the benefit of battery approaches and standardization to behaviour studies and identify study design parameters and criteria that determine study reproducibility.

Materials and Methods

Animal care

All animal experiments conformed to the British Home Office Regulations (Animal Scientific Procedures Act 1986) and local ethical approval, with specific procedures carried out under Wellcome Trust Sanger Institute guidelines for care and use of animals. Animals were housed in individually vented cages (Seal Safe 1284 L; Tecniplast, Italy) with 5 animals per cage, enriched with paper tunnels, on a normal 12-hour light, 12 hour dark cycle. Animals were provided with IPS 5021 Autoclavable Mouse Breeder Diet (LabDiet, St Louis, MO, USA) and water ad libitum.

Mouse generation.

The present study focused on three loss-of-function or knockout mutations: 1) a deletion of essential exons of the *Dlg4* gene, which encodes the PSD95 protein (Migaud *et al.*, 1998); 2) a deletion of exons of the *Dlg3* gene, which encodes the SAP102 protein (Cuthbert *et al.*, 2007); and 3) a deletion of exons of the *Dusp6* gene (manuscript in preparation). All were generated by gene targeting of 129 embryonic stem cells and backcrossing onto both 129S5/SvEvBrd (129S5) background (PSD95 ≥ 99.97% isogenic, median 99.98%; SAP102 ≥ 98.9% isogenic, median 99.87%; DUSP6 ≥ 97.9% isogenic, median 97.9%) and C57BL/6J background (PSD95 ≥ 75.0% isogenic, median 87.5%; SAP102 ≥ 87.5 % isogenic, median 87.5%; DUSP6 ≥ 99.2% isogenic, median 99.2%) in the heterozygous state. Intercrossing generated cohorts of mice for behavioural measurements. All mutations were made in-house. Wildtype 129S5/SvEvBrd and C57BL/6J mice used for back-crossing were obtained from stocks maintained at the Wellcome Trust Sanger Institute.

For comparisons of behaviour in wildtype mice, 46 129S5 and 51 C57BL/6J mice were used, achieving >90% power to detect a phenotype of Cohen *d* = 0.7 at α = 0.05. In mutant versus wildtype comparisons, group sizes were 10-20 mutants and 10-20 wildtypes, corresponding to 56.7-87.5% power to detect phenotypes of Cohen *d* = 1 at α = 0.05.

Within a line, mutant and wildtype mouse cohorts were litter-matched, sex-matched and age-matched, but across background strains, age matching was not attempted since mutation effect sizes were insensitive to observed age effects (see Results). Age, sex and numbers of mice are shown in Table 1.

Behavioural testing.

A one-week standardised protocol was applied consisting of five test environments, with each animal assayed in all environments in the same order. The experimenter was blinded to sex and genotype using radio-frequency identifying chips implanted before the behaviour experiments. Experimenter bias was eliminated by measuring behaviour variables automatically using cameras and other devices.

Day 1, elevated plus maze (EPM). The EPM had two exposed arms and two walled arms 45 cm above the ground, with infrared illumination and monitored by an IR-filtered digital camera (all from Tracksys, Nottingham, UK), connected to Mediacruise software, version 2.24.000. The elevated plus maze had arms 10 cm wide and 60 cm long. The central zone

was defined as the 10 cm by 10 cm area in the centre of the maze. Mice were placed on the open arm next to the central zone. The body centroid of the mouse was used as the mouse location. Analysis was carried out using Noldus Ethovision software, version 3.1.16 (Tracksys, Nottingham, UK). Mice were run on the maze for five minutes in a darkened room under red light, two mazes per room, with the experimenter in the room. After each run, mazes were cleaned with ethanol wipes.

Day 2, morning, open field (OF). The OF was a white, matte-finish plastic 75 cm by 75 cm box with 42 cm walls on an infrared bed monitored by a digital camera (all available from Tracksys, Nottingham, UK). Animals explored the box for five minutes under normal room lighting (1000 lux), and behaviour was analysed by Noldus Ethovision software. After each run, the apparatus was cleaned with ethanol wipes.

Day 2, afternoon, novel object exposure (NOE). The NOE assay used the OF apparatus, with an unopened aluminium 355ml soft drink can in the centre of the box. The animal explored the apparatus in the presence of the novel object for five minutes. Mice were assayed in OF and NOE in the same order, and the interval between OF and NOE was approximately four hours.

Day 3, rotarod (RR). An accelerating rotarod apparatus with 3.0 cm diameter spindle 35 cm above the bottom of the apparatus was used (EZ-ROD, version 2.12, Accusan Instruments, Columbus, Ohio, USA). The spindle accelerated from 10 revolutions per minute (RPM) to 48 RPM over five minutes. A mouse's fall triggered a computer-monitored switch, and latency to fall and maximum spindle speed were recorded for each trial. Each mouse underwent eight trials in the morning and eight in the afternoon. Within a session, trials began approximately 30 seconds after the end of the previous trial. After each mouse, the apparatus was cleaned with ethanol wipes.

Days 4-5, classical conditioning. Training was done in an operant box system from Coulbourn Instruments (Whitehall, PA, USA). After two minutes of habituation, a 300 Hz tone at 83-86 dB was played for 30 seconds, co-terminating with a 2-second scrambled shock in the grid floor at 0.45 mA under control of Acctimetrics FreezeFrame software. Two more tone-shock pairings were presented at 100-second intervals. Mouse behaviour was recorded by overhead video camera and freezing behaviour was detected by Acctimetrics

FreezeView software, version 2. Testing was performed 24 hours later in the same boxes. After three minutes' exposure to the operant box, the tone was played for two minutes, and freezing was recorded in 30-second time bins.

Identification of 16 maximally independent behaviour variables

All manipulation of behaviour variables was done in the R statistical programming environment. One hundred and five highly redundant raw behaviour variables were summarised by a reduced set of 16 variables. Candidate behaviour variables were principally hand-selected for their own historical interest or relatedness to behaviours of historical interest. The resulting list of variables was trimmed based on two statistical criteria: 1) independence from (lack of correlation with) one another, and 2) wildtype similarity across 129S5 and C57BL/6J background, which was assessed by performing Chi-squared goodness of fit tests of the (scaled and shifted) z-score distribution of each variable in each background.

Five variables described behaviour in EPM: 1) EPM total distance (cm); 2) EPM max speed (cm/s); 3) EPM % time in open; 4) EPM time in centre (s); and 5) EPM max speed, open vs closed (cm/s).

OF and NOE contributed two variables. 1) Total distance (cm) travelled during the open field and novel object exposure assays was denoted 'OF, NOE total distance'. Log_{10} transformation of these scores increased similarity of score distributions between 129S5 and C57BL/6J mice and resulted in approximately normally-distributed data. 2) Response to the change in environment from the OF to NOE was measured by the difference between distances travelled (cm) between the two assays, denoted 'NOE vs OF distance travelled'.

RR measured a mouse's innate motor coordination, motor learning, and motor memory, which were derived as shown in Figure S1A. Two linear models were fit, one to a mouse's latency to fall during the eight trials in the morning session, and the other to the eight trials in the afternoon session. Naïve performance, denoted 'RR naive fall time', was computed as the fitted value of motor performance in the second trial in the morning session. Motor learning, denoted 'RR learning', was measured as the slope of the linear model during the morning session. Motor memory, denoted 'RR memory', was measured as the difference between the fitted midpoint of the afternoon session and the fitted midpoint of the morning

session. This model of naive performance, learning, and memory had the following properties across all G2C wildtype mice (n_{129S5} = 851, $n_{C57BL/6J}$ = 395): 1) learning is not correlated with naive performance; and 2) the measure of memory is positively correlated with the measure of learning and more modestly with the measure of naive performance. RR naive fall time was $\log_{10}$ transformed to increase similarity between the 129S5 and C57BL/6J score distributions. This transformation resulted in approximately normally-distributed data.

Classical conditioning contributed six variables related to learning and memory. The derivation of the variables is depicted in Figure S1B-D. We noted that the two learning (memory acquisition) variables were not strictly independent. Not only did they reference the same raw data, we hypothesised that increases in the tone response might in part be driven by general increases in freezing due to contextual learning alone. To detect separable aspects of learning, all G2C data for 129S5 or C57BL/6J mice were used to construct a linear regression model relating the tone effect (Learning, tone effect; LRN_tone) to the general increase in freezing in successive trials (Learning, trial effect; LRN_trial); linear dependence between the two was subtracted from LRN_tone. Sample R code for this operation is:

```
LRN_tone = lm(LRN_tone ~ LRN_trial)$residuals
```

This was done separately for mice on the C57BL/6J background and on the 129S5 background.

Similar to learning variables, memory variables were interdependent. Cued memory responses, ‘Cued memory, mean’ (CU_mean) and ‘Cued memory, change’ (CU_change), were not independent of the contextual response, (Contextual memory, mean; CT_mean). Furthermore, the temporal evolution of the cued response (Cued memory, change; CU_change) was expected to be correlated with the mean cued response, (Cued memory, mean; CU_mean). To derive a measure of the temporal change in the cued response independent of the context effect and the mean cued effect, linear dependencies of CU_change on CT_mean and CU_mean were subtracted, again using linear models. This was done separately for C57BL/6J mice and 129S5 mice. Sample R code for this operation is:

```
CU_change = lm(CU_change ~ CT_mean + CU_mean)$residuals
```

Dependence of the mean cued effect, CU_mean, on the context effect, CT_mean, was then subtracted similarly, using R code:

```
CU_mean = lm(CU_mean ~ CT_mean)$residuals
```

Steps taken to reduce data interdependence amongst classical conditioning behaviour variables conferred the following properties to classical conditioning variables: 1) the trial effect during task acquisition (Learning, trial effect; LRN_trial) is now exclusively predictive of the contextual memory variable (Contextual memory, mean; CT_mean); and 2) the tone effect during task acquisition (Learning, tone effect; LRN_tone) is now exclusively predictive of the cued memory effect (Cued memory, mean; CU_mean). In other words, the mathematical operations presented here discover separable aspects of task acquisition that separately predict two aspects of memory. A caveat, however, is that these operations are only robust when based on data from many mice, whereas similar calculations on cohort sizes such as 20 mutants and 20 wildtypes are likely to suffer from over-fitting.

The final behavioural repertoire consisted of 16 minimally-redundant/minimally-correlated behavioural variables, of which eight represented innate, instinctive behaviours and eight involved learning and memory.

Statistical analysis of mutant-wildtype differences

The mice and experiments reported here were generated as part of a pipeline operation and were tested in weekly batches. This design permitted robust assessment and control of the effects of four confounding factors: batch effect, two aspects of ageing, and variance differences between mutant and wildtypes. The batch effect arises as a result of differences in experimental conditions from week to week. Ageing may affect both mutants and wildtypes, or one of the groups only, increasing the variance of one or both groups and thus decreasing power to detect a phenotype. Although mice were litter-matched and thus age-matched within mouse lines, substantial age differences were present in our dataset, ranging from approximately six weeks to over one year old. Furthermore, median ages differed between mouse lines, necessitating analysis of ageing. In addition to batch and age, we analysed variance differences between mutant and wildtype animals, which may

erode the significance of mutant-wildtype differences. We performed these analyses on simulated behavior data, which permitted specification of true effect sizes. Subsequently, comparison of trends in experimental data to trends in simulated data permitted approximation of the magnitudes of confounding effects. By simulating the effect of these magnitudes of confounding effects on simple two-factor (mutant genotype, and sex if appropriate) ANOVAs, we could rule out any major impact of confounding effects. Therefore significance of phenotypes was assessed by ANOVA with respect to sex and genotype at the mutant locus. These analyses were performed in the R statistical programming environment are detailed in Supporting Text 1.

Effect size calculations for single behaviour variables and groups of variables.

Phenotypic magnitude was computed as the maximum likelihood estimator of Cohen *d* effect size, which for balanced experimental designs is approximately equal to the standardised mean difference of two groups (Cohen, 1988). Aggregate effect sizes of multiple variables were computed by averaging effect size magnitudes; this measure is by definition always positive or minimally zero. Therefore, for each aggregate effect size, random draws of wildtype data of the same genetic background were used to simulate null effect; the median of this distribution was taken as the baseline null effect. Combined effect sizes for mutant lines of mice were required to significantly exceed this baseline to be detected as significant.

Standard error of effect size was computed by sampling 1000 times from wildtype mice matched with the experimental cohorts for cohort size, genetic background and sex. This strategy makes no assumption about the distribution of the underlying data and therefore is more accurate than estimation of the standard error of effect size using a canonical formula.

Comparison of mutant phenotypic effect size across genetic backgrounds.

One of the main concerns of this work was to measure the likelihood of phenotypic recapitulation (that is, that phenotype effect sizes did not significantly differ) across backgrounds. For a given phenotype comparison, defined by a given mutation and

behaviour variable, effect size difference between 129S5 and C57BL/6J was tested for significance using a z-score computed as the difference of *d* divided by the combined standard error: $Z = -|d_2 - d_1|/(SE_1^2+SE_2^2)^{0.5}$. The two-tailed p-value was assessed from the standard normal distribution.

Impact of sampling and p value threshold on expected reproducibility

We modelled the effect of sampling and the commonly used $p<0.05$ significance threshold on phenotypic reproducibility under the assumption of identical experimental conditions. This analysis is detailed in Supporting Text 2.

Results

Construction and validation of a standardised behavioural dataset

The G2C programme characterised behaviour of 60 mutant mouse lines (10-20 litter-matched mice per group) in a standardised behaviour pipeline. Fifty-five of these lines were on either 129S5 or C57BL/6J background; six lines (Table 1) reflect three mutations (PSD95/*Dlg4*, SAP102/*Dlg3*, and *Dusp6*) introduced into both these backgrounds. Of 105 raw behaviour variables measured per mouse, 16 variables efficiently summarised the dataset; these were minimally-correlated and thus minimally-redundant (Figure S2). Amongst wildtype mice of the same background (n_{129S5} = 851, $n_{C57BL/6J}$ = 395) these variables showed Pearson $R^2 < 0.016$. Mechanistic relationships existed between 9/120 pairs of variables (learning versus memory of a task, and variables assessing aspects of locomotion); these showed median Pearson $R^2 = 0.15$. Wildtype z-score transformed behaviour scores were similarly distributed for both backgrounds for 14/16 variables (Chi-squared goodness of fit tests, $p > 0.05$). The strain z-score distributions differed only in that 129S5 mice showed a broader range of times in the EPM open arm and C57BL/6J mice showed a broader range of RR memory. These results indicate the sixteen variables were minimally redundant and similar between wildtype mice in the two background strains.

We next validated our assays by comparison with previously published results. As shown in Figure S3 and Figure 1, comparisons of wildtype mice from the two background strains recapitulated known behaviour differences between 129 and C57BL/6 mice. Consistent with previous reports, 129S5 mice showed lower activity in EPM total distance, EPM max speed,

OF NOE total distance (Bolivar *et al.*, 2000, Rodgers *et al.*, 2002); reduced motoric performance on the rotating rod task, with males more severely impacted than females (Mcfadyen *et al.*, 2003); more freezing in response to contextual fear; and increased occupancy of the central zone of the elevated plus maze (Bolivar *et al.*, 2001, Rodgers *et al.*, 2002). We noted two discrepancies with published reports. First, in our experiments 129S5 mice spent longer in the open arm of the EPM (Rodgers *et al.*, 2002), which is most likely explained by the protocol's minimally-sized central zone and the relative inactivity of 129S5 mice (which were placed just outside the central zone). Second, 129S5 mice did not differ from C57BL/6Js in mean freezing in our cued fear memory assay, in contrast to a published report (Bolivar *et al.*, 2001), in which C57BL/6 had higher activity indices than 129 mice. This difference may have resulted from differences in the protocol. These results establish that our behaviour protocols recapitulate known behaviour patterns in these strains.

Minimal impact of four confounding effects.

Prior to analysis of the impact of the mutations, it was necessary to examine potential confounding effects. Our pipelined study design, which used litter matched mice, was subject to four such effects: batch, unequal variance, age and age × genotype. Effects like these have been shown to play a role in phenotyping pipelines and to be amenable to analysis using mixed effect models (Karp *et al.*, 2012). We used a similar mixed effect model approach, as described in Supporting Text 1, which showed that confounding effects generally eroded the effect sizes, but this erosion was usually less than 20% overall. Therefore analyses of variance were restricted to sex and genotype at the mutant locus.

Two contrasting methods of assessing similarity/reproducibility of mutant phenotypes across experiments

Although it is widely recognised that reproducibility of phenotypes between studies performed on different strains is important, to our knowledge there have been no methods available to measure this. The first and most elegant way to address this question is quantitative and compares the effect sizes of phenotypes. In this case, we expect that two published results, one significant and the other not, may yet be insignificantly different in effect size and therefore be considered as equivalent. Assuming a test-retest scenario with

no methodological variation between the published reports, effect sizes are expected to differ at the expected false positive rate, usually 5%.

A second method of assessing reproducibility refers to phenotypic p-values and addresses the situation found in phenotype databases derived from literature reports. Such databases record presence or absence of a phenotype, indicating that the null hypothesis test yielded a p-value less than the accepted false positive rate, usually 5%. Thus, this method of comparing studies accepts phenotypes that reach $p<0.05$ in either of the two published experiments and asks how many phenotypes found in one report are found in the other. Assuming a test-retest scenario with no methodological or strain variation between the published reports, simulations showed that with a sample size of 15 animals per group, sampling variation places an upper bound of 60.6% reproducibility on such pairs of experiments (see Materials and Methods). It should be noted that with further repetitions the effect of sampling will be overcome and the likelihood of detecting a significant phenotype will increase above 60.6%.

Reproducibility of phenotype effect sizes.

One of the main objectives of this work was to measure the likelihood that the effect of a mutation would be recapitulated across backgrounds. To assess this likelihood, we compared mutant phenotypes across mouse genetic backgrounds in terms of their effect size, rather than in terms of real-world units (cm and seconds), which may vary considerably between strains. This strategy involves scaling of phenotypes separately by background strain. Examination of the 16 behavioural variables in the PSD95, SAP102 and DUSP6 mutant mice on both backgrounds revealed 41 of 48 tests (85%) showed no significant difference between background strains (Figure S4, summarised in Figure 2). Plotting effect sizes in C57BL/6J background against effect sizes in 129S5 background revealed a strong correlation (16 variables × three mutations, Spearman rho = 0.489, p = 0.0005, data not shown). Thus, both backgrounds revealed overall similarity in both innate and learned behaviours.

Because the ability to measure phenotypic differences is affected by sampling, we next compared our cohorts of 10-20 mutants to large datasets of background-matched wildtype mice from 55 G2C lines (n_{129S5} = 851, $n_{C57BL/6J}$ = 395, except SAP102, for which $n_{129S5,male}$ =

451, $n_{C57BL/6J,male}$ = 215; data not shown). With the larger and more robust wildtype dataset we found that the number of detected strain differences increased from seven to seventeen. These differences occurred with similar frequency in all three mutations (seven, four, and six differences in PSD95, SAP102, and DUSP6 lines, respectively). This observation is reinforced by the fact that the measured effect size differences (48 comparisons) had a standard deviation (0.53 Cohen *d* units) similar to the sizes of standard errors of the phenotypic effect size estimates (96 individual measures, median standard error 0.53 Cohen *d* units). Masking of strain differences by sampling effects implies that, for the broad behaviour repertoire presented here, sampling effects associated with sample sizes typically used in mutant-wildtype comparisons mask most strain effects. Thus strain effects are not a necessary barrier to comparison of mutant behaviour phenotypes between mouse backgrounds.

One of the advantages of our behaviour battery is the opportunity to measure magnitude and direction of mutant phenotypes for a large number of semi-independent measures of mouse behaviour (semi-independence shown in Figure S2). However, it is also of interest to develop aggregate measures of phenotypic magnitude/severity. Mutations with restricted effects will have smaller aggregate size than mutations with broad effects on multiple behaviour variables. It should be noted that these measures do not summarise a mutant phenotypic profile, but merely measure in aggregate how extreme it is.

We obtained such aggregate phenotypic measures by averaging absolute values of Cohen *d* (litter matched mutant-wildtype experiments, Figure 3) across three sets of behaviour variables (all 16 variables, eight innate, and eight learned). A beneficial property of these aggregate phenotypes is that they are subject to less sampling variation, having in effect a lower “signal to noise ratio” than the component phenotypes. This is similar to the manner in which averaging the results of several tests gives a more robust measure of a school pupil’s performance in a given subject. Another beneficial property of averaging may be cancellation of some background strain-related variation, which may permit reliable ranking of genes by their overall phenotypic impacts in mice irrespective of genetic background (and thus be a key measure of phenotype that might have translational relevance).

As shown in Figure 3A-C, using litter-matched wildtypes, averaging all 16 phenotypes (Figure 3A, Overall) showed a strain difference between DUSP6 lines. Averaging the eight

innate behaviours (Figure 3B) showed no strain differences and averaging the eight learned behaviours (Figure 3C) showed a difference between strains in PSD95 mutants. We next asked if these differences could be sampling artefacts, and therefore recalculated all combined behaviour effect sizes based on the larger set of (background-matched, Figure 3D-F) wildtype mice. This analysis showed no phenotypic differences in PSD95 and DUSP6 mice and a nominally significant difference in learned behaviour in SAP102 mutants. These results show that aggregate phenotypes are similar in both genetic backgrounds. They also suggest that sampling has a larger impact than strain background on these measures.

Reproducibility of phenotypic p-values.

As noted above, phenotype databases derived from literature reports record presence or absence of phenotypes passing a p-value threshold. Applying this criterion to our dataset of three mutations, 16 behavioural variables, and two backgrounds, we found that only 36% (10/28) of phenotypes were reproduced across mouse genetic backgrounds, consistent with a combination of strain and sampling effects. This is less than the 17 phenotypes (60.6%) expected when there is no strain difference ($p = 0.007$, binomial test; supplementary analysis detailed in Supporting Text 2).

We next sought to identify a more stringent significance threshold for phenotypes that would result in a greater than 36% reproducibility in another strain background. We scanned a broad range of increasingly stringent p-value thresholds in each background and counted behaviour phenotypes that achieved nominal significance in the same direction in the other background (see “this” and “other” strain in Figure 4). As expected, increases in p-value stringency of observed phenotypes led to an increasing fraction of nominally-significant other-strain phenotypes. However, it should be noted that increasing p-value stringency and thus increasing reproducibility of phenotypes concomitantly results in fewer phenotypes being detected and thus increased risk of false negative phenotypes.

For 48 behaviour tests, the optimal fraction of reproduced phenotypes was 50-60% and this was achieved in the range $0.0011 < p < 0.01$ (see window in Figure 4). At maximum, 60% (6/10) phenotypes were reproduced at a phenotype threshold of $p = 0.0024$ (Figure 4). This test was repeated using the larger cohorts of background-matched wildtypes: within the

same threshold window 56-68% phenotypes were reproduced (data not shown). Thus, reproducibility of significant mutant phenotypes is increased by the use of larger sets of background-matched controls.

Discussion

We have used behavioural data from a large and standardised phenotyping programme to characterise the effect of 129S5 or C57BL/6J genetic background on behaviour of mice carrying single targeted mutations in three different genes. Whilst this is not the first report of a modulating effect of genetic background on mutant mouse behaviour, the extant reports are rare, anecdotal, variable in methods, and affected by sampling variation and other confounds, and therefore cannot offer an estimate of the prevalence of this phenomenon. Besides being systematic, the importance of this study is two-fold. First, it addressed substrains of the most commonly used mouse strains in behavioural research, which differ greatly in wildtype behaviour. Second, it addresses behaviour measures that we expect to affect mouse performance in many other commonly used behaviour assays.

Recapitulation of mutant phenotypes across different genetic backgrounds was assessed by measurements of 16 variables, describing broad aspects of innate and learned behaviour, in five apparatuses. The validity of the majority of the protocols is reinforced by the fact that behaviours of wildtype mice in our experiments recapitulated the results of others. Three mutations afforded a total of 48 comparisons, which provided an estimate of recapitulation frequency across these background strains. With our mutant-wildtype group sizes, 85% of mutant phenotypes had indistinguishable effect sizes between 129S5 and C57BL/6J; overall the effect sizes were also highly correlated. Importantly, this showed that large wildtype differences do not necessarily translate into large impacts on mutant phenotype effect size. Rather, the mutant phenotypes in one strain were merely a scaled version of the phenotypes in the other, leading to typically similar effect sizes.

Composite measures of function are used in many areas, for example in human IQ and mental health screening (Xu *et al.*, 2015). Whereas in human measures, summary measures are made at the level of the individual, in our experiments summary measures are made at the level of the mutant mouse line. Averaging the absolute mutant effect sizes of multiple individual variables produced a composite estimate of overall mutant effect.

Applying this approach, we found very similar overall impacts of mutation on both backgrounds. We propose that aggregate mutant phenotypes like these represent new and useful forms of behaviour measurement. For example, this measure could serve as a quantitative score for the overall impairment of the behavioural repertoire in lines of mutant mice and to identify genes (or drugs) of greatest overall importance in behaviour.

A challenge faced by any study that addresses the reproducibility of phenotypes is to develop criteria for equivalence with published results. Literature and public databases describe mutant behavioural phenotypes in terms of presence (nominal significance) and direction and therefore equivalence criteria should be described in these same terms. Using the significance threshold of $p<0.05$, which is typically applied in behavioural research, we found that only 36% of phenotypes were reproduced across strains compared to an expected maximum of 60.6% in the absence of strain and methodological variation. It should be noted that these levels of reproducibility were obtained under a highly standardized behavioural testing and housing protocol and would be expected to be lower between laboratories or with other methodological heterogeneity.

To search for an optimum in reproducibility, we used a sliding p-value threshold approach and found that 50% or more of mutant phenotypes were reproducible when the original p-value was $p<0.01$. The observation that less stringent thresholds resulted in less reproducible results is relevant for interpreting published behaviour results and for the design of experiments. Specifically, this p-value threshold approach, in which greater reproducibility is measured by nominal significance upon retest in any background strain, profits from higher experimental power and larger group sizes. We show that phenotype effect sizes observed at larger group sizes differ between mouse strains more frequently. However, for practical group sizes of 10-20 animals, we also show that strain-based mutant phenotype differences are masked by sampling effects. Furthermore, we find evidence that at these modest group sizes, averaging across multiple phenotypic variables permits robust assessment of phenotypic severity of mouse mutations irrespective of background strain. This experimental robustness is achieved by standardised behavioural testing in a test battery that probes a wide range of behaviours.

Eight caveats apply to generalising the results of our study. First, phenotypic reproducibility between studies is more likely in the presence of methodological homogeneity in assay

content and analysis methods; in practice, it is on this level that many studies differ. Second, recapitulation frequency is subject to sampling variation, and thus partitioning of behaviours into those that do or do not recapitulate across strains will not be definitive in our study; this was demonstrated by the fact that significant strain differences in our aggregate measures of behaviour were abolished by analysis of a much larger wildtype dataset. A third caveat is that our high frequency of recapitulation was observed in a broad behavioural repertoire; conceivably, assays focusing on narrower aspects of cognition may exhibit an altered rate of recapitulation. Fourth, when comparing results from two backgrounds, it may be necessary to scale measures to achieve relevance across background strains. Fifth, it should be observed that the robustness of aggregate behaviour measures depends on having multiple measures, some of which, by themselves, may not achieve nominal significance. Sixth, a technical limitation was that C57BL/6J isogenicity was low in two of the mouse lines; in spite of this, expected behavioural differences between 129S5 and C57BL/6J wildtypes were confirmed. Seventh, we acknowledge that the number of mutations in this study was low, raising the question of overall generalisation of our observation of similarity in mutant behaviour profiles in future studies. What this study has demonstrated, however, is that when comparing mutations in two behaviourally very different strains a large modulatory effect of background strain on mutant phenotypes is not necessarily to be expected. Importantly, this study also suggests that inherent uncertainty in mutant-wildtype studies due to sampling effects may be more important than genetic background in comparisons between studies. An eighth caveat, that there is more than one way to measure frequency of modulation of mutant phenotypes by genetic background, is illustrated by a new report by Sittig *et al* (2016). Those authors described frequent strain × mutation interactions based on analysis of known phenotypes of two genes in heterozygous F1 crosses. It should be noted that focusing on known phenotypes inflates the prior likelihood of strain differences. Furthermore, where strong background effects are rare, Sittig *et al*'s use of one ANOVA per phenotype (across many strains) further inflates the relative frequency with which significant strain × mutation interactions are identified. On the other hand, comparing minimally correlated phenotypes using effect sizes (as we did) would reveal a lower estimate of frequency of strain × mutation interactions.

In conclusion, our study validates the use of this rapid and simple behaviour test battery for characterising and comparing mutant mice with respect to a broad repertoire of innate and

learned behaviours. It also suggests the value of aggregate behaviour measures for comparison of overall phenotype severity between mouse lines and genetic backgrounds.

References

Baker, M. (2011) Animal models: inside the minds of mice and men. *Nature,* **475,** 123-128.

Balogh, S.A., McDowell, C.S., Stavnezer, A.J. & Denenberg, V.H. (1999) A behavioral and neuroanatomical assessment of an inbred substrain of 129 mice with behavioral comparisons to C57BL/6J mice. *Brain Res,* **836,** 38-48.

Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A. & Richardson, J.E. (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res,* **42,** D810-817.

Bolivar, V.J., Caldarone, B.J., Reilly, A.A. & Flaherty, L. (2000) Habituation of activity in an open field: A survey of inbred strains and F1 hybrids. *Behav Genet,* **30,** 285-293.

Bolivar, V.J., Pooler, O. & Flaherty, L. (2001) Inbred strain variation in contextual and cued fear conditioning behavior. *Mamm Genome,* **12,** 651-656.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences,* L. Erlbaum Associates.

Crabbe, J.C., Wahlsten, D. & Dudek, B.C. (1999) Genetics of mouse behavior: interactions with laboratory environment. *Science,* **284,** 1670-1672.

Crawley, J.N. (2008) Behavioral phenotyping strategies for mutant mice. *Neuron,* **57,** 809-818.

Crawley, J.N., Belknap, J.K., Collins, A., Crabbe, J.C., Frankel, W., Henderson, N., Hitzemann, R.J., Maxson, S.C., Miner, L.L., Silva, A.J., Wehner, J.M., Wynshaw-Boris, A. & Paylor, R. (1997) Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. *Psychopharmacology (Berl),* **132,** 107-124.

Cuthbert, P.C., Stanford, L.E., Coba, M.P., Ainge, J.A., Fink, A.E., Opazo, P., Delgado, J.Y., Komiyama, N.H., O'Dell, T.J. & Grant, S.G. (2007) Synapse-associated protein 102/dlgh3 couples the NMDA receptor to specific plasticity pathways and learning strategies. *J Neurosci,* **27,** 2673-2682.

Drapeau, E., Dorr, N.P., Elder, G.A. & Buxbaum, J.D. (2014) Absence of strong strain effects in behavioral analyses of Shank3-deficient mice. *Dis Model Mech,* **7,** 667-681.

Gerlai, R. (2001) Gene targeting: technical confounds and potential solutions in behavioral brain research. *Behav Brain Res,* **125,** 13-21.

Grant, S.G., O'Dell, T.J., Karl, K.A., Stein, P.L., Soriano, P. & Kandel, E.R. (1992) Impaired long-term potentiation, spatial learning, and hippocampal development in fyn mutant mice. *Science,* **258,** 1903-1910.

Grant, S.G.N. (2003) An integrative neuroscience program linking mouse genes to cognition and disease. In Plomin, R., DeFries, J.C., Craig, I.W. & McGuffin, P. (eds), *Behavioral genetics in the postgenomic era.* American Psychological Association, Washington, DC, US, pp. 123-138.

Holmes, A., Lit, Q., Murphy, D.L., Gold, E. & Crawley, J.N. (2003) Abnormal anxiety-related behavior in serotonin transporter null mutant mice: the influence of genetic background. *Genes Brain Behav,* **2,** 365-380.

Holmes, A., Wrenn, C.C., Harris, A.P., Thayer, K.E. & Crawley, J.N. (2002) Behavioral profiles of inbred strains on novel olfactory, spatial and emotional tests for reference memory in mice. *Genes Brain Behav,* **1,** 55-69.

Huang, H.S., Burns, A.J., Nonneman, R.J., Baker, L.K., Riddick, N.V., Nikolova, V.D., Riday, T.T., Yashiro, K., Philpot, B.D. & Moy, S.S. (2013) Behavioral deficits in an Angelman syndrome model: effects of genetic background and age. *Behav Brain Res,* **243,** 79-90.

Karp, N.A., Melvin, D. & Mott, R.F. (2012) Robust and sensitive analysis of mouse knockout phenotypes. *PLoS One,* **7,** e52410.

Kastenberger, I., Lutsch, C., Herzog, H. & Schwarzer, C. (2012) Influence of sex and genetic background on anxiety-related and stress-induced behaviour of prodynorphin-deficient mice. *PLoS One,* **7,** e34251.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F., Cuthill, I.C., Fry, D., Hutton, J. & Altman, D.G. (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One,* **4,** e7824.

Logue, S.F., Paylor, R. & Wehner, J.M. (1997) Hippocampal lesions cause learning deficits in inbred mice in the Morris water maze and conditioned-fear task. *Behavioral neuroscience,* **111,** 104-113.

McFadyen, M.P., Kusek, G., Bolivar, V.J. & Flaherty, L. (2003) Differences among eight inbred strains of mice in motor ability and motor learning on a rotorod. *Genes Brain Behav,* **2,** 214-219.

Migaud, M., Charlesworth, P., Dempster, M., Webster, L.C., Watabe, A.M., Makhinson, M., He, Y., Ramsay, M.F., Morris, R.G., Morrison, J.H., O'Dell, T.J. & Grant, S.G. (1998) Enhanced long-term potentiation and impaired learning in mice with mutant postsynaptic density-95 protein. *Nature,* **396,** 433-439.

Mistry, A.M., Thompson, C.H., Miller, A.R., Vanoye, C.G., George, A.L., Jr. & Kearney, J.A. (2014) Strain- and age-dependent hippocampal neuron sodium currents correlate with epilepsy severity in Dravet syndrome mice. *Neurobiology of disease,* **65,** 1-11.

Morice, E., Denis, C., Giros, B. & Nosten-Bertrand, M. (2004) Phenotypic expression of the targeted null-mutation in the dopamine transporter gene varies as a function of the genetic background. *Eur J Neurosci,* **20,** 120-126.

Nithianantharajah, J., Komiyama, N.H., McKechanie, A., Johnstone, M., Blackwood, D.H., St Clair, D., Emes, R.D., van de Lagemaat, L.N., Saksida, L.M., Bussey, T.J. & Grant, S.G. (2013) Synaptic scaffold evolution generated components of vertebrate cognitive complexity. *Nat Neurosci,* **16,** 16-24.

Park, S.S., Stranahan, A.M., Chadwick, W., Zhou, Y., Wang, L., Martin, B., Becker, K.G. & Maudsley, S. (2011) Cortical gene transcription response patterns to water maze training in aged mice. *BMC neuroscience,* **12,** 63.

Pietropaolo, S., Guilleminot, A., Martin, B., D'Amato, F.R. & Crusio, W.E. (2011) Genetic-background modulation of core and variable autistic-like symptoms in Fmr1 knock-out mice. *PLoS One,* **6,** e17073.

Popova, N.K., Naumenko, V.S., Tibeikina, M.A. & Kulikov, A.V. (2009) Serotonin transporter, 5-HT1A receptor, and behavior in DBA/2J mice in comparison with four inbred mouse strains. *J Neurosci Res,* **87,** 3649-3657.

Rodgers, R.J., Boullier, E., Chatzimichalaki, P., Cooper, G.D. & Shorten, A. (2002) Contrasting phenotypes of C57BL/6JOlaHsd, 129S2/SvHsd and 129/SvEv mice in two exploration-based tests of anxiety-related behaviour. *Physiol Behav,* **77,** 301-310.

Sittig, L.J., Carbonetto, P., Engel, K.A., Krauss, K.S., Barrios-Camacho, C.M. & Palmer, A.A. (2016) Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. *Neuron,* **91,** 1253-1259.

Xu, M.K., Gaysina, D., Barnett, J.H., Scoriels, L., van de Lagemaat, L.N., Wong, A., Richards, M., Croudace, T.J., Jones, P.B. & Group, L.H.A.G. (2015) Psychometric precision in phenotype definition is a useful step in molecular genetic investigation of psychiatric disorders. *Translational psychiatry,* **5,** e593.

Acknowledgments

Financial support is acknowledged from the Wellcome Trust, the UK Medical Research Council, and the European Union 7th Framework Programmes EUROSPIN (FP7-HEALTH-241498), SynSys (FP7-HEALTH-242167) and GENCODYS (FP7-HEALTH-241995). We thank members of the Grant laboratory for helpful discussion, as well as Mike Allerhand and Ian Deary for statistics advice. The authors declare no conflict of interest.

Figures and Legends

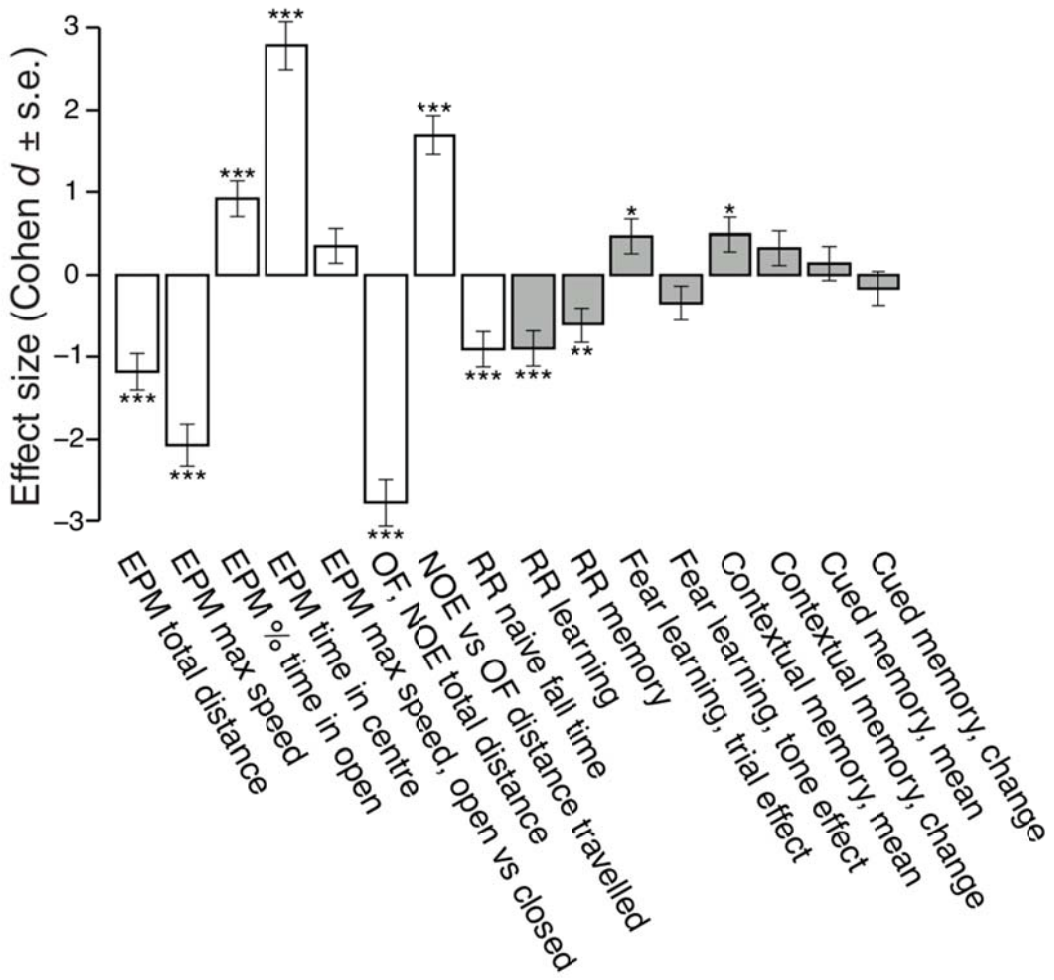

Figure 1 – Main effects of genetic background on behaviour.

Bar plot depicts effect sizes (Cohen *d,* 129S5 versus C57BL/6J) of behavioural differences in Figure S3, Supporting information. White bars, innate behaviours; shaded, learned behaviours. Error bars (SE) are computed using the standard formula based on sample sizes and magnitude of the effect size. Significance of difference between strains shown: *p<0.05, **p<0.01, ***p<0.001. Apparatus: EPM, elevated plus maze; OF, open field; NOE, novel object exposure; RR, rotating rod. Individual variables described in Table 2.

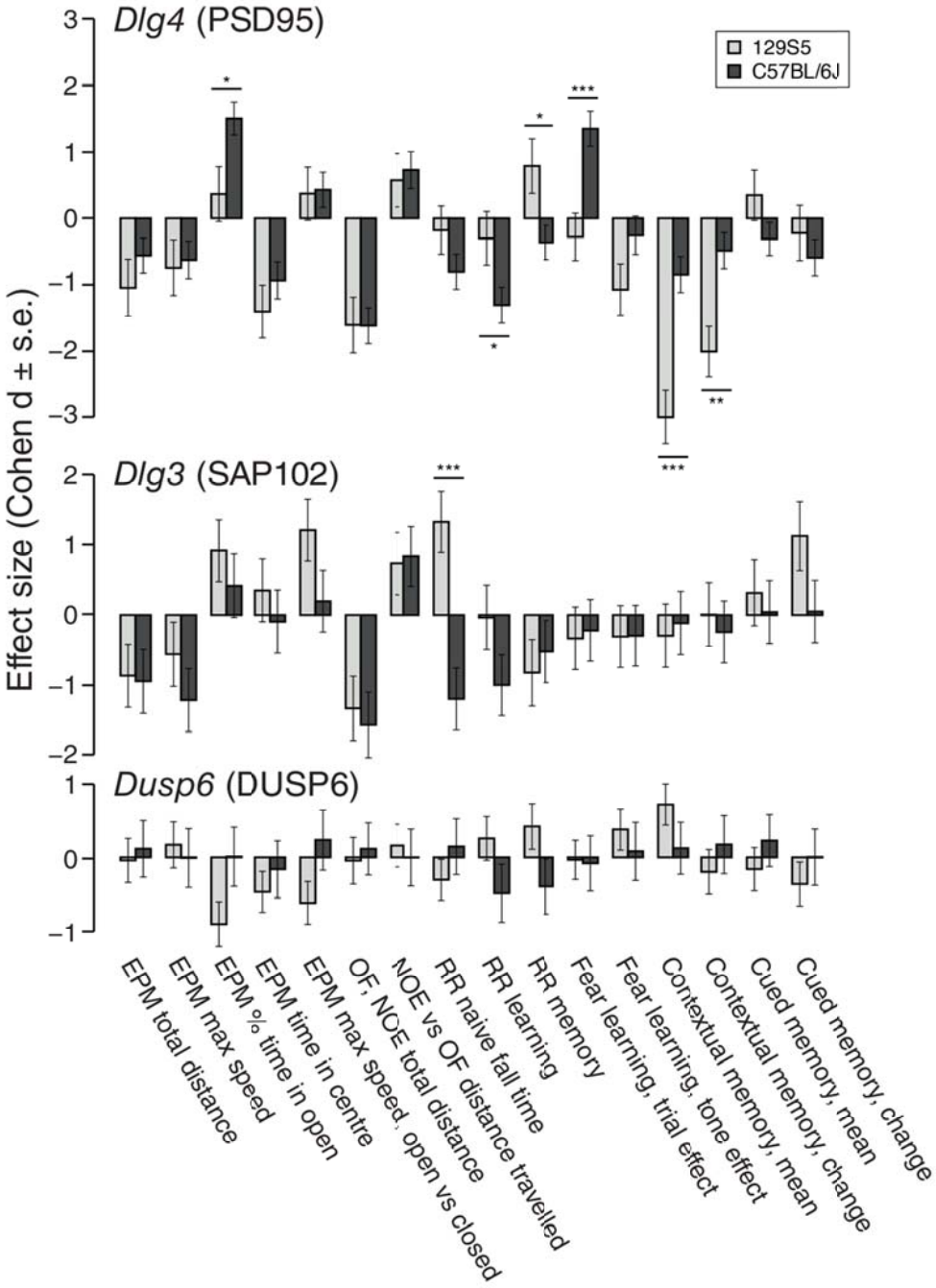

Figure 2 – Effects of genetic background on mutant behavioural phenotypes.

Comparison of phenotypic outcome (Cohen *d* ± SE, mutant versus wildtype) in 16 behavioural variables on 129S5 (lighter bars) and C57BL/6J (darker bars) genetic backgrounds. Z-test significance: *p<0.05, **p<0.01, ***p<0.001. Related behaviour profiles are found in Figure S4, Supporting information.

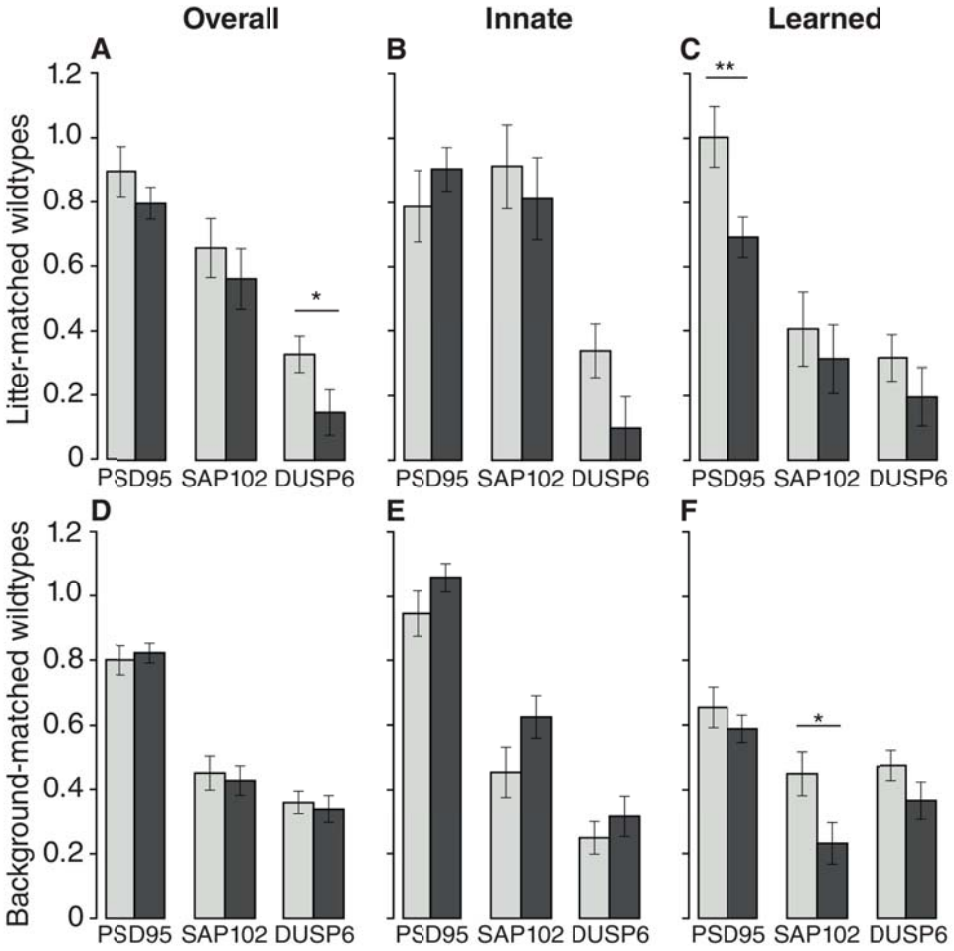

Figure 3 – Aggregate effect sizes of three mutations in two background strains.

Combined effect sizes calculated from litter-matched mutants and wildtypes (A-C) and from comparisons of mutants with background-matched wildtypes (D-F)(n_{129S5} = 851; $n_{C57BL/6J}$ = 395; except SAP102, for which $n_{129S5,male}$ = 451, $n_{C57BL/6J,male}$ = 215).

(A,D) Overall effect sizes (average of both innate and learned).

(B,E) Combined Innate effect sizes (average of 8 innate variables).

(C, F) Combined Learned effect sizes (average of 8 learning/memory variables).

Lighter bars, 129S5; darker bars, C57BL/6J.

Error bars, SE; Z-test significance: *p<0.05, **p<0.01.

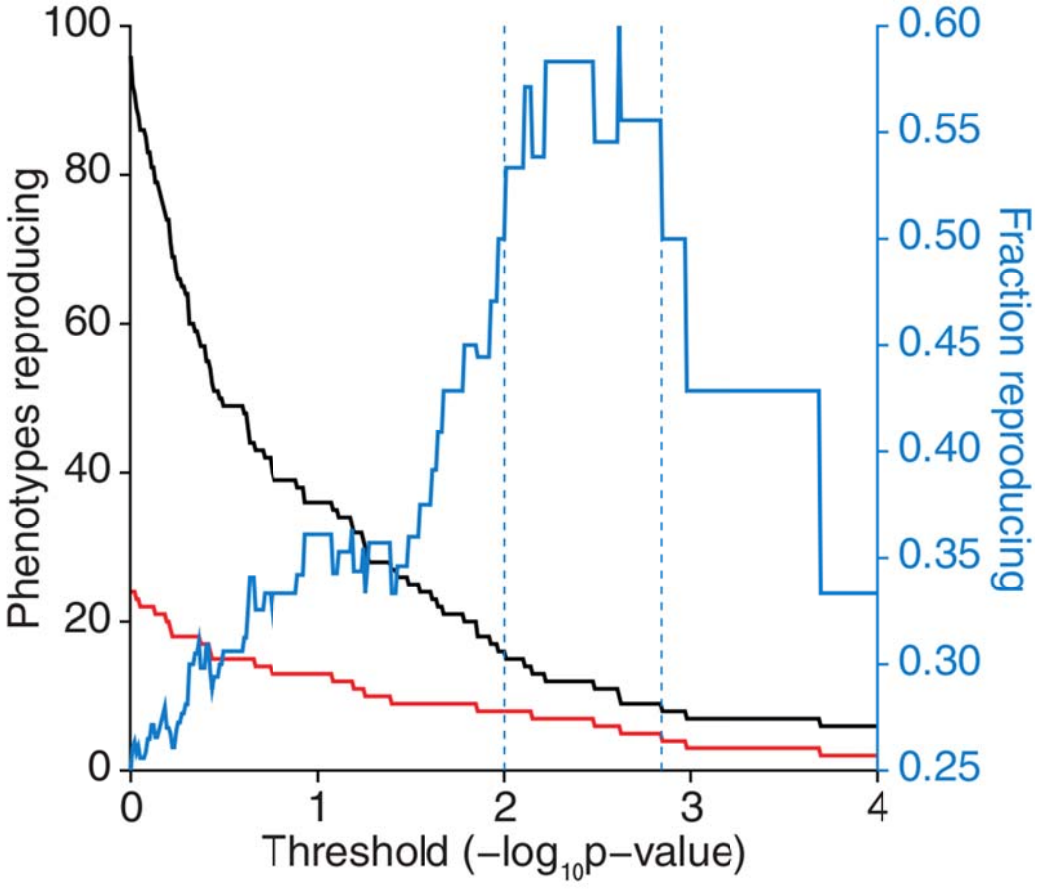

Figure 4 – Reproducibility of mutant phenotypes between strains.

The number (maximum 96, left axis) of phenotypes above threshold in “this” strain (black line) that reproduce with $p < 0.05$ in the “other” strain (red line) is shown. Threshold stringency increases from left to right. The fraction of phenotypes that reproduce (blue line, right axis) generally increases with greater stringency of the threshold. Dotted lines span the window where >50% of phenotypes reproduce.

Tables

Table 1: numbers, sex, and age of mice

Mutation	Genetic background	Wildtypes (F,M)	Mutants (F,M)	Age (days, quantile: min, 0.25, 0.75, max)
PSD95/*Dlg4*	129S5	6,8	6,7	40,65,88,194
	C57BL/6J	11,16	13,16	56,136,207,285
SAP102/*Dlg3*	129S5	0,11	0,11	183,224,300,322
	C57BL/6J	0,10	0,14	48,121,208,329
DUSP6/*Dusp6*	129S5	10,11	11,13	48,164,248,294
	C57BL/6J	7,7	9,7	107,249,383,472

Table 2: Innate and learned behaviour variables and definitions

Assay	Variable name	Type of variable	Description
Elevated plus maze (5 min)	EPM total distance	innate	Total distance (cm) travelled in arm or central zone
	EPM max speed	innate	Maximum speed in any arm or central zone (cm/s)
	EPM % time in open	innate	% time spent in open arms, per total time spent in arms
	EPM time in centre	innate	Time (s) spent in the central zone, not in the arms
	EPM max speed open vs. closed	innate	Difference in maximum speed (cm/s) in the open arm versus closed arm
Open field & novel object exposure (5 min each)	OF, NOE total distance	innate	Total distance (cm) travelled in two assays, $\log_{10}$ transformed
	NOE vs. OF distance travelled	innate	Difference in distance travelled (cm) in novel object exposure versus open field
Rotating rod	RR naïve fall time	innate	Second time to fall from the rotating rod (s) during eight trials in the morning, from fitted linear model, $\log_{10}$ transformed
	RR learning	learned	Rate of increase of fall time per trial (s/trial) in the morning session, from fitted linear model
	RR memory	learned	Difference in fall time (s) between midpoint performances in afternoon and morning, from fitted linear models
Classical conditioning	Learning, trial effect	learned	Increase in % freezing during third pair of stimuli versus first

Assay	Variable name	Type of variable	Description
training/acquisition			pairing
	Learning, tone effect	learned	Increase in tone response (% freezing) due to third tone versus tone response at first tone
Classical conditioning context memory	Contextual memory, mean	learned	Mean % freezing during two min of context re-exposure versus initial two min of habituation during training
	Contextual memory, change	learned	Change in % freezing during context re-exposure
Classical conditioning cue memory	Cued memory, mean	learned	Mean increase in % freezing due to tone
	Cued memory, change	learned	Change in % freezing in last versus first time bin during audible cue